Learning Multisensory Representations Using Predictive Coding

Parva Alavian (s.p.alavian@uva.nl)
Cognitive and Systems Neuroscience Group, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, Netherlands

Kwangjun Lee (k.lee@uva.nl)
Cognitive and Systems Neuroscience Group, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, Netherlands

Matthias Brucklacher (m.m.brucklacher@uva.nl)
Cognitive and Systems Neuroscience Group, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, Netherlands

Jorge Mejias (j.f.mejias@uva.nl)
Cognitive and Systems Neuroscience Group, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, Netherlands

Sander Bohte (sbohte@cwi.nl)
Cognitive and Systems Neuroscience Group, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, Netherlands
Machine Learning Group, Centrum Wiskunde & Informatica, Amsterdam, Netherlands

Cyriel Pennartz (c.m.a.pennartz@uva.nl)
Cognitive and Systems Neuroscience Group, Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, Netherlands

Abstract

Integration of information arriving in the brain from different sensory modalities is essential for robust perception. In this work, we use the predictive coding framework - a prominent theory of cortical processing- to perform multisensory representation learning. Our model can learn meaningful joint representations from two separate streams of data. These representations function as a form of hetero-associative memory, allowing the network to recall or reconstruct one modality from the other. The reconstructed outputs preserve class-relevant features, even in the absence of one sensory modality. These results suggest that predictive coding networks can serve as a biologically plausible framework for modeling multisensory representation learning.

Keywords: Multisensory integration; Predictive coding; Multisensory perception; Representation learning

Introduction

Multisensory integration (MSI) refers to the brain's ability to combine information from multiple sensory modalities. MSI plays a critical role in perception by allowing us to form meaningful representations of the sensorily rich environment around us. This integration is essential for effective action planning and for resolving perceptual ambiguities and uncertainties (Ernst & Bülthoff, 2004). Despite its importance, the underlying neural mechanisms of MSI remain largely unknown.

Predictive coding, an increasingly influential theory of cortical processing, has been proposed as a unifying framework for perception, action, and cognition (Friston, 2005; Pennartz et al., 2019; Rao & Ballard, 1999). According to this framework, the brain continuously generates predictions about incoming sensory stimuli, compares these predictions with actual inputs, and uses the resulting prediction errors to learn better predictions. (Rao & Ballard, 1999) applied these principles to build a hierarchical three-layer predictive coding network that mimicked neuronal responses in the visual cortex. Various extensions to predictive coding models have been proposed thereafter (Spratling, 2017), but most focus on a single modality. Notable exceptions include (Pearson et al., 2021) which explored visuo-tactile integration for place recognition in robots. Despite the essential role that MSI plays in perception, predictive coding models that address multisensory inference and cross modal predictions remain scarce.

In this work, we extend the classical predictive coding network of Rao and Ballard to perform multisensory inference. We demonstrate that such a network is capable of learning multisensory representations of data and using these representations to perform cross-modal recall.

Model and training procedure

Figure 1 presents a schematic of our model, which consists of three modules. Two unisensory streams represent two distinct, arbitrary sensory modalities (both visual in these experiments for simplicity). Each stream follows the Rao and Ballard architecture, with each layer in the network consisting of two groups of neurons: representation neurons that predict the activity of the layer below and error neurons that project the mismatch between the prediction and the actual representation to the layer above.

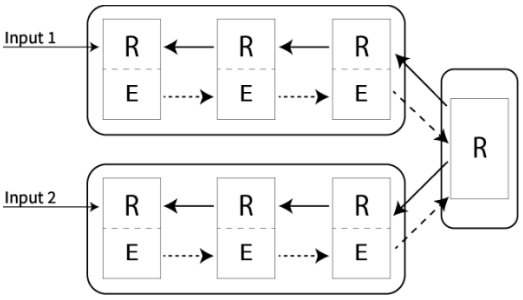

Figure 1: Model schematic. R and E denote populations of representation and error neurons, respectively. Solid lines indicate predictions; dotted lines indicate errors.

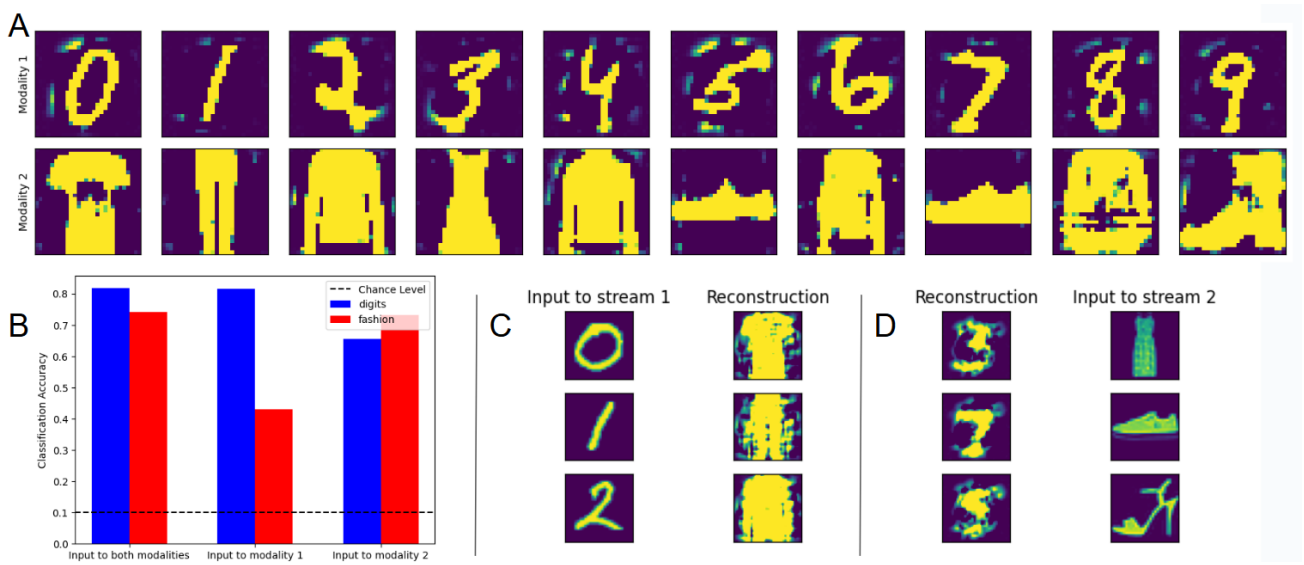

Figure 2: Panel A shows pairs of reconstructed images from the test dataset. Panel B shows the classification accuracy of generated images. Panels C and D show cross-modal reconstructions of a missing input using the available modality.

At the highest level, the two unisensory streams converge into a multisensory module. This module jointly predicts the activity of the topmost layers of both unisensory pathways and receives prediction errors from both modalities.

We train the network on paired inputs from the MNIST and Fashion-MNIST datasets. Modality 1 receives digit images from MNIST, while modality 2 is exposed to fashion item images from Fashion-MNIST. The datasets are paired by class—for example, digit "1" in modality 1 is always paired with the fashion item "pants" in modality 2, though the specific samples vary across instances.

Results

Our findings demonstrate that the proposed multisensory network is capable of learning meaningful multimodal representations. Figure 2, Panel A shows reconstructed images of the test data, obtained by decoding the representations inferred at the joint multisensory layer through the feedback pathway of the network. These reconstructions—generated via top-down predictions—preserve class-relevant features, as evidenced by Figure 1, Panel B, where a logistic regression model—trained on the original MNIST and Fashion-MNIST datasets—successfully classifies the generated images.

Cross-Modal Recall. A key advantage of multisensory processing is robustness in the face of partial sensory input—such as when one modality is missing or noisier than the other. Our model demonstrates this capability by reconstructing samples from the missing modality using only the available input from the other modality. Figure 2, Panels C and D illustrate reconstructed images when one modality -fashion items and digits respectively- is absent. Figure 2, Panel B shows that these reconstructions are classifiable by a logistic regression model trained on the original dataset, confirming that essential features are preserved even in the absence of one modality.

Conclusions

We have shown that a multisensory network based on predictive coding can learn robust multisensory representations and perform cross-modal recall. Our findings provide insights into the computational mechanisms underlying multisensory perception and highlight the potential of predictive coding-based models in cognitive and artificial intelligence applications.

References

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. In *Trends in Cognitive Sciences* (Vol. 8, pp. 162–169). https://doi.org/10.1016/j.tics.2004.02.002

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Pearson, M. J., Dora, S., Struckmeier, O., Knowles, T. C., Mitchinson, B., Tiwari, K., Kyrki, V., Bohte, S., & Pennartz, C. M. A. (2021). Multimodal Representation Learning for Place Recognition Using Deep Hebbian Predictive Coding. *Frontiers in Robotics and AI*, *8*. https://doi.org/10.3389/frobt.2021.732023

Pennartz, C. M. A., Dora, S., Muckli, L., & Lorteije, J. A. M. (2019). Towards a Unified View on Pathways and Functions of Neural Recurrent Processing. In *Trends in Neurosciences* (Vol. 42, Issue 9, pp. 589–603). Elsevier Ltd. https://doi.org/10.1016/j.tins.2019.07.005

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

Spratling, M. W. (2017). A review of predictive coding algorithms. In *Brain and Cognition* (Vol. 112, pp. 92–97). Academic Press Inc. https://doi.org/10.1016/j.bandc.2015.11.003